University of Warsaw  
Faculty of Economic Sciences

Bartłomiej Kowalczuk

Album no. 372926

Michał ThorAlbum no. 361309

**Predicting bankruptcy among Polish companies using logistic regression**

Paper for Advanced Econometrics

subject during Data Science studies

Warsaw, may 2019

INTRODUCTION

Bankruptcies among companies were and still are widely discussed because of their impact on the economy. In fact bankruptcies have both good and bad influence on the macroeconomic situation – on one hand its’ occurrences result in loss of jobs and create market uncertainty[[1]](#footnote-1) (especially when bigger company defaults) but on the other hand it is an instrument of clearing the market from redundant firms.[[2]](#footnote-2) Although it may be useful, most of the market participants would rather know when certain company might “go down”. Focused efforts by economists and statisticians, that began in the thirties, effected in variety of tools designed to work as a precautionary mechanisms. However, prediction accuracy of these tools, especially econometric models, is often questioned. Considering Poland, most of the approaches from literature[[3]](#footnote-3) used to predict company bankruptcy are based on the discriminant analysis (such as Altman’s model) and logistic regression (such as Ohlson “O-score”). Also data used in those examples is not sampled – it is mainly based on finding every existing bankrupt company in the population and randomly attaching same number of healthy companies.

Main goal of our paper is to determine whether it is possible to create prediction model that gives reasonable company bankruptcy predictions on unbalanced data (in which bankrupt companies make only 3% of the whole dataset – as it is far more realistic scenario) in great advance (5 years prior to bankruptcy) using logistic regression. Afterwards we try to compare it to model 1 year prior to default. If models appear to be applicable then we want to examine which accounting variables are the main determinants of these bankruptcies (and whether these determinants differ between models).

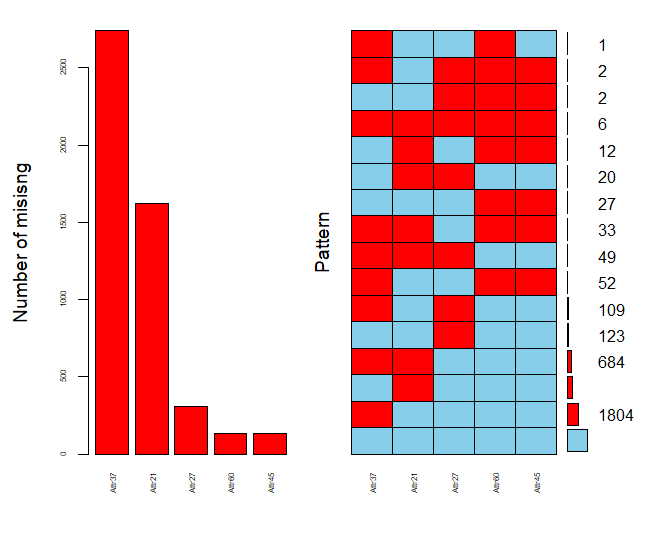
SECTION I

Preparing the dataset

Data which we used comes from a polish bankruptcies dataset on UCI Machine Learning Repository.[[4]](#footnote-4) Firstly we analyzed data for 1st year of forecasting period and therefore defaults mean that company default 5 years from that point. All of the basic 64 variables are transformed accounting data and various financial indicators (e.g. RoA, RoE, etc.). There were initially 7027 observations including 271 defaults which make for ~3.5% of the dataset.

Firstly we analyzed how many missing data are there in the set. Some variables were more prone to have NA’s but overall 3833 observations were not complete (including at least one missing variable). Number of missing for 5 most uncomplete variables are seen on the Figure 1.

Figure 1 - Variables with the most missing data

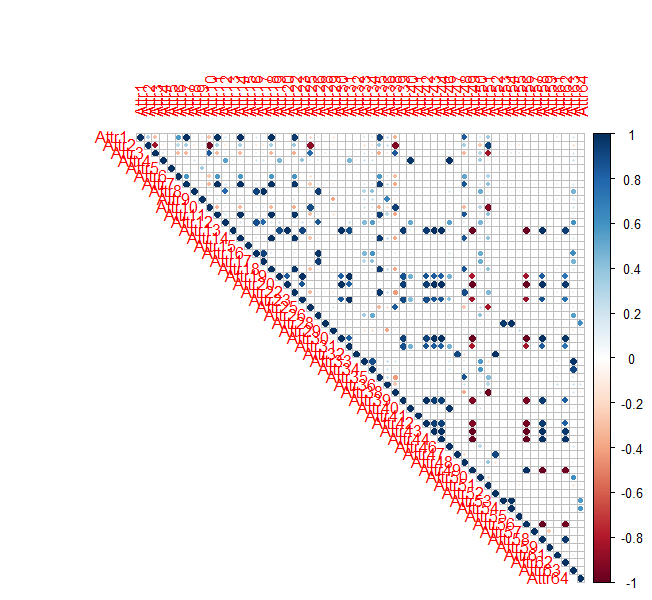


Wanting to retain most possible information while removing variables with unacceptable ratio of missing data, we introduced 5 factor variables indicating that there were missing data:

Subsequently we omitted columns with over 3% missing observations from further analysis. Since most of the missing came from the same observations we removed all of the (around 200 observations). Having in mind that retaining as much defaults as possible is crucial we find out that there are still around 3.5% of defaults. For some of the observations for which we already had the factor variable we decided to replace NA with median from the dataset (because they were single cases).

Next we analyzed correlation between variables. Correlation matrix can be seen on Figure 2.

Figure 2 - Correlation matrix for variables before removals

**

It can be seen that there are a lot of highly correlated variables. We filtered all the variables that are correlated with at least 2 other with correlation higher than 90% . Unfortunately it resulted in half of the variables being gone. Fortunately though it meant that there were almost no high correlated data and no missing observations at all. Correlation matrix for the cleaned data can be seen on Figure 3.

Figure 3 - Correlation matrix after removals

Obraz zawierający tekst, mapa

Opis wygenerowany automatycznie

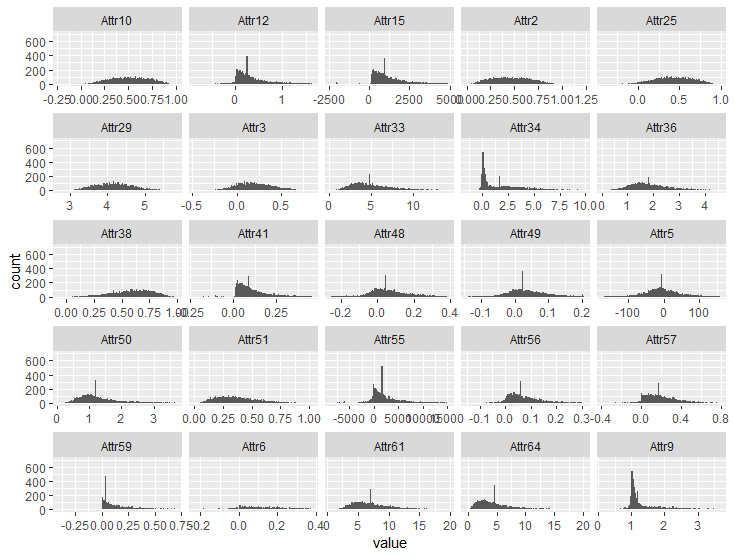
Correlations look much better, although some of them are still high. As we will be trying to retain as much variables as possible we will add those correlated variables as interactions when building model.

Further problem lies in the distribution of variables – there are A LOT of outliers in the data. With first estimations combined problems of high correlation and very high variance of most of the variables resulted in complete separation and Hauck Donner effect – instabilitiy of estimation in logisitc regression. To confront that problem we decided to replace some of the outliers with other values:

where: IQR – interquartile range

Therefore we are left with 4254 observations, while still retaining same percantage of defaults in dataset. On Figure 4 histograms of all (beside factor) variables (in Attr59 and Attr6 there are missing bins for zeros as they were larger than the scale – we shall emphasize on it below) are shown.

Figure 4 - Histograms for all variables



We can see that data is far from normally distributed and there are still values on the edges of distribution graph, but the improvement is undeniable. Also as mentioned below, Attr6 and Attr59 both have around 45% of zeros within its values but we decided to keep them in furhter analysis – maybe those with positive value of retained earnings/long-term liabilities are more prone to default.

Next step in our analysis is to transform some variables. After computing Information Values (IV) for all of the remaining variables, we decided that Attr12 should be binned as its suspiciously high predictive power (of 0.9) would work better when binned into two groups with borderline value of 0.06 which splits variable into groups with default rates of 14% and 2% respectively.

We also added interactions between correlated variables, interactions that would give us nominal values of accounting measures and some transformations to variables which had very high IQR.

SECTION II

Estimating models

With cleaned and prepared dataset with almost the same percentage of defaults as initial dataset we started estimating first, basic logarithmic regression model. At first we split our dataset into training and testing sets with a 70/30 proportion retaining default percentage in both subgroups. All of the models were validated using repeated k-fold cross validation for estimation of forecasts with 5 folds and 5 repeats.

First model was estimated on all variables with interactions. A lot of variables were statistically insignificant and so the AIC (Akaike’s Information Criterion) was 514.74. We set the cut-off point on 0.02, just to be sure that we are identifying as much bankrupts as we can. Its’ accuracy on test data was 79.76% but sensitivity (which in predicting bankruptcies should be our main goal) was only 83.72%. Estimated forecast sensitivity and specificity were 75.6% and 79.3% respectively.

Second model was estimated using same variables but we applied forward and back propagation (simultanously doing general-to-specific and specific-to-general methods) minimizing AIC as our priority (if removing certain variable improves AIC – remove it). We were left with 20 variables, although not all of them statistically significant at p-value = 0.05. AIC was 485.58 and estimated forecasts were looking much better than in the first model (Sens. = 82.4% and Spec. = 76.7%), however accuracy of the model and its’ sensitivity on test data did not improve – area under ROC curve and accuracy fell down.

At last we tried removing all those interactions and estimating model on only cleaned and prepared variables without interactions and transformed variables using AIC “both-side propagation” mentioned in second model. That way we obtained third model with AIC = 500.91 – worse than in second model. Estimated accuracy was similar to the second model but its’ area under ROC on test data for cut-off point of 0.02 improved to 91.78% and, more importantly, sensitivity increased to 86.05%.

Comparison of the models’ coefficients estimations can be seen in the Appendix 1.

Comparison of the models’ accuracy is presented in Table 1 – AUC means area under curve. Underlined are the best values in each metric. Despite not being the most accurate overall we chose Model 3 as the best one because of the highest sensitivity and simpler form with less variables.

Table 1 - Comparison of models' performance on test data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **AIC** | **Accuracy** | **Sensitivity** | **Specificity** | **AUC** |
| Model 1 | 514.74 | 79.76% | 83.72% | 79.63% | 91.00% |
| Model 2 | 485.58 | 78.04% | 83.72% | 77.84% | 90.77% |
| Model 3 | 500.91 | 75.14% | 86.05% | 74.76% | 91.78% |
| Model 3.1 | 642.83 | 65.73% | 86.04% | 65.02% | 87.93% |

Considering Model 3 as our best model, we decided to remove all the statistically insignificant variables () so the odds ratios and marginal effects would be interpretable and significant – resulting in Model 3.1.

In Model 3.1 we were left with only 9 variables but the drop in accuracy was visible. In case of interpretability we decided that because sensitivity estimations and sensitivity on test data did not differ much between Model 3 and 3.1 to chose Model 3.1 as our final estimation. We then conducted the Hosmer and Lemeshow (H-L) GoF (Goodness of Fit) test to determine whether our model fits the data well. As we eliminated overfitting from our analysis (because estimated forecasts were similar to the ones obtained on test data) we overcame main flaw of the H-L test. P-value of the test = 0.29 indicates that there is indeed good fit and event rates among 10 groups match expected rates.

In Table 2 we present determinants along with their odds ratio – impact on the odds of bankruptcy. means that for one unit increase of a variable the odds of defaulting are times lower. For we know that odds of defaulting increase by (or by ). For we know that that varaible does not have an impact on odds of going bankrupt.

Table 2 - Odds ratios and descriptions of variables in Model 3.1

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Odds ratio** |
| Attr6 | retained earnings /  total assets | 0.003 |
| Attr12 | gross profit /  short-term liabilities | 0.323 |
| Attr33 | operating expenses /  short-term liabilities | 1.136 |
| Attr38 | constant capital /  total assets | 0.032 |
| Attr49 | EBITDA /  sales | 991.234 |
| Attr57 | (curr. assets – inventory – short-term liabilities) /  (sales – gross profit – depreciation) | 0.223 |
| Attr651 | no long-term liabilities | 0.073 |
| Attr661 | no financial expenses | 75.527 |

It can be seen that only Attr33, Attr49 and Attr66 impact odds of bankruptcy positively (increase odds of bankrupting) – despite the level of the magnitude which is not interpretable due to specificity of the dataset. It is also seen in the Table 3 where marginal effects for Model 3.1 are provided.

Table 3 - Marginal effects for Model 3.1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **AME** | **SE** | **z** | **p** | **lower** | **upper** |
| Attr12 | -0.0278 | 0.0124 | -2.2390 | 0.0252 | -0.0522 | -0.0035 |
| Attr33 | 0.0031 | 0.0012 | 2.5172 | 0.0118 | 0.0007 | 0.0056 |
| Attr38 | -0.0844 | 0.0191 | -4.4117 | 0.0000 | -0.1219 | -0.0469 |
| Attr49 | 0.1698 | 0.0641 | 2.6481 | 0.0081 | 0.0441 | 0.2955 |
| Attr57 | -0.0369 | 0.0194 | -1.9068 | 0.0565 | -0.0749 | 0.0010 |
| Attr6 | -0.1410 | 0.0386 | -3.6535 | 0.0003 | -0.2167 | -0.0654 |
| Attr651 | -0.0144 | 0.0058 | -2.4864 | 0.0129 | -0.0257 | -0.0030 |
| Attr661 | 0.4827 | 0.0520 | 9.2809 | 0.0000 | 0.3808 | 0.5847 |

AME columns gives us information on how, with unchanged other values, change of that particular variable would influence the bankruptcy probability. Therefore it is visible, that not having any financial expenses increases probability of default by . On the other hand, increase of ratio by 0.1 would decrease probability of default by 1.41 percentage points. All (but Attr57) variables are statistically significant, therefore interpretable.

However, that set of significant variables in Model 3.1 is slightly different than in literature[[5]](#footnote-5) on bankruptcy prediction among polish companies as there is no sign of financial indicators such as RoA, WCR or RoE (*Return on Assets, Working Capital Ratio and Return on Equity*). One element of financial statement is reoccurring in both literature and our model – short-term liabilities. It might be that those models were estimated using data for 1 year prior and our model is estimated 5 years prior to bankruptcy so there lies the difference.

We were curious if this difference is reproducible and would be the same in our analysis – therefore we conducted second estimation on data 1 year prior to the bankruptcy. That dataset consisted of 5910 observations with 410 defaults.

Having in mind that the paper should not be too long and readable, we omitted data preparation and feature selection description – we followed similar approach as with the first dataset so we will just show the model estimation and the results for comparison.

BIBLIOGRAPHY

Carter R., Van Auken H.: *Small Firm Bankrupcy*.Journal of Small Business Management 2006 44(4), 2006, pages 493–512

Kiesielińska, J., Waszkowski A.: *Polish models to predict bankruptcy and its verification,* Ekonomika i organizacja gospodarki żywnościowej 82, Wydawnictwo SGGW, 2010, pages 17-31

Mączyńska E., Zawadzki M.: *Dyskryminacyjne modele predykcji bankructwa przedsiębiorstw*. Ekonomista 2/2006.

Ptak-Chmielewska A.: *Prediction models of SME bankruptcy in Poland – analysis using Cox survival model and logistic regression model*. Ekonometria 4 (46), 2014, pages 9-21

APPENDICES

Appendix 1 - Table of coefficients for Models 1,2,3

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Results** | | | | | |
|  | | Dependent variable:  class | | | |
|  | Model 1 | | Model 2 | Model 3 | Model 3.1 |
| Attr2 | 13.747 | | 16.449\*\*\* | 5.252\*\* |  |
|  | (12.023) | | (6.130) | (2.196) |  |
| Attr3 | 6.280 | |  | -1.931\* |  |
|  | (13.339) | |  | (1.064) |  |
| Attr5 | 0.002 | |  |  |  |
|  | (0.005) | |  |  |  |
| Attr6 | -3.443 | | -4.405\*\* | -3.458\* | -5.730\*\*\* |
|  | (2.248) | | (1.891) | (1.877) | (1.505) |
| Attr9 | 0.614 | |  | 0.386\* |  |
|  | (0.436) | |  | (0.211) |  |
| Attr10 | 3.439 | | 5.208\*\* | 4.008\* |  |
|  | (2.882) | | (2.378) | (2.358) |  |
| Attr12 |  | |  | -1.384\*\* | -1.131\*\* |
|  |  | |  | (0.668) | (0.502) |
| Attr15 | 0.0003\* | | 0.0003\*\* | 0.0002\* |  |
|  | (0.0001) | | (0.0001) | (0.0001) |  |
| Attr25 | -2.468 | |  |  |  |
|  | (3.628) | |  |  |  |
| Attr29 | 0.542 | |  |  |  |
|  | (1.032) | |  |  |  |
| Attr33 | 0.233\*\* | | 0.174\*\* | 0.095 | 0.127\*\* |
|  | (0.102) | | (0.068) | (0.065) | (0.050) |
| Attr34 | 0.124 | |  |  |  |
|  | (0.110) | |  |  |  |
| Attr36 | -0.582 | |  |  |  |
|  | (0.363) | |  |  |  |
| Attr38 | -6.445\*\* | | -6.460\*\*\* | -7.035\*\*\* | -3.429\*\*\* |
|  | (2.787) | | (2.096) | (2.059) | (0.737) |
| Attr41 | 0.399 | |  |  |  |
|  | (1.869) | |  |  |  |
| Attr48 | 1.501 | |  |  |  |
|  | (2.534) | |  |  |  |
| Attr49 | 0.523 | |  | 13.182\*\*\* | 6.899\*\*\* |
|  | (5.439) | |  | (3.251) | (2.565) |
| Attr50 | 0.588 | | 0.805\* | 0.734\* |  |
|  | (0.498) | | (0.444) | (0.396) |  |
| Attr51 | 10.517 | |  | -4.909\*\*\* |  |
|  | (8.720) | |  | (1.823) |  |
| Attr55 | -0.001\* | | -0.001 |  |  |
|  | (0.001) | | (0.001) |  |  |
| Attr56 | -9.862\*\*\* | | -7.761\*\*\* | -6.295\*\* |  |
|  | (2.996) | | (2.590) | (2.464) |  |
| Attr57 | -2.839\*\* | | -2.925\*\*\* | -3.180\*\*\* | -1.500\* |
|  | (1.173) | | (0.938) | (1.066) | (0.780) |
| Attr59 | 0.504 | |  |  |  |
|  | (0.980) | |  |  |  |
| Attr61 | 0.017 | |  |  |  |
|  | (0.160) | |  |  |  |
| Attr64 | -0.175 | | -0.067 |  |  |
|  | (0.153) | | (0.050) |  |  |
| Attr651 | -1.536\*\*\* | | -1.292\*\*\* | -0.851\*\* | -2.623\*\*\* |
|  | (0.452) | | (0.385) | (0.349) | (0.400) |
| Attr661 | 13.491\*\*\* | | 6.052\*\*\* | 5.740\*\*\* | 4.324\*\*\* |
|  | (4.376) | | (0.429) | (0.393) | (0.303) |
| Attr671 | 0.647 | |  |  |  |
|  | (0.990) | |  |  |  |
| Attr681 | 22.791 | | 21.413 | 22.145 |  |
|  | (678.988) | | (728.862) | (670.732) |  |
| Attr12\_BINNED | 0.350 | |  |  |  |
|  | (0.405) | |  |  |  |
| I(Attr22) | 0.512 | | -6.331 |  |  |
|  | (8.316) | | (4.367) |  |  |
| I(Attr253) | 9.247\*\* | | 9.285\*\* |  |  |
|  | (4.513) | | (3.656) |  |  |
| I(log1p(Attr61)) | -0.437 | |  |  |  |
|  | (1.421) | |  |  |  |
| I(log1p(Attr64)) | 0.848 | |  |  |  |
|  | (1.308) | |  |  |  |
| Attr2:Attr29 | -0.326 | |  |  |  |
|  | (1.372) | |  |  |  |
| Attr661:Attr12\_BINNED | 0.579 | |  |  |  |
|  | (0.875) | |  |  |  |
| Attr38:Attr661 | -6.905 | |  |  |  |
|  | (4.407) | |  |  |  |
| Attr9:I(exp(Attr29)) | -0.001 | |  |  |  |
|  | (0.004) | |  |  |  |
| Attr2:Attr51 | -17.778\* | | -5.018\*\* |  |  |
|  | (10.777) | | (2.240) |  |  |
| Attr2:Attr661 | -6.253 | |  |  |  |
|  | (4.023) | |  |  |  |
| Attr2:Attr3 | -4.578 | |  |  |  |
|  | (13.802) | |  |  |  |
| Attr3:Attr10 | 11.018 | | 14.307\*\* |  |  |
|  | (13.591) | | (7.275) |  |  |
| Attr3:Attr25 | -4.844 | |  |  |  |
|  | (5.074) | |  |  |  |
| Attr3:Attr38 | -13.151 | | -13.189\*\* |  |  |
|  | (9.531) | | (6.593) |  |  |
| Attr25:Attr38 | 0.253 | | -6.054\*\* |  |  |
|  | (7.983) | | (2.660) |  |  |
| Attr49:Attr56 | 93.948\*\*\* | | 93.214\*\*\* |  |  |
|  | (26.153) | | (19.227) |  |  |
| Attr29:Attr55 | 0.0002\* | | 0.0002 |  |  |
|  | (0.0001) | | (0.0001) |  |  |
| Constant | -12.625\* | | -9.171\*\*\* | -3.952\* |  |
|  | (6.861) | | (3.282) | (2.311) |  |
| Observations | 2,979 | | 2,979 | 2,979 | 2,979 |
| Log Likelihood | -210.369 | | -219.791 | -232.456 | -312.042 |
| Akaike Inf. Crit. | 514.738 | | 485.583 | 500.912 | 642.083 |
| Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | | | | | |

1. Carter R., Van Auken H.: *Small Firm Bankrupcy*.Journal of Small Business Management 2006 44(4), 2006, pages 493–512 [↑](#footnote-ref-1)
2. Mączyńska E., Zawadzki M.: *Dyskryminacyjne modele predykcji bankructwa przedsiębiorstw*. Ekonomista 2/2006. [↑](#footnote-ref-2)
3. Kiesielińska, J., Waszkowski A., *Polish models to predict bankruptcy and its verification,* Ekonomika I organizacja gospodarki żwynościowej 82, 2010, Wydawnictwo SGGW, pages 17-31 [↑](#footnote-ref-3)
4. Accessible on: [*https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data*](https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data) [↑](#footnote-ref-4)
5. Ptak-Chmielewska A.: *Prediction models of SME bankruptcy in Poland – analysis using Cox survival model and logistic regression model*. Ekonometria 4 (46), 2014, s. 9-21 [↑](#footnote-ref-5)